ARTICLES

Journal of Semiconductors
(2024) 45, 012301
doi: 10.1088/1674-4926/45/1/012301

Optimized operation scheme of flash-memory-based neural network online training with ultra-high endurance

Yang Feng[1], Zhaohui Sun[1], Yueran Qi[1], Xuepeng Zhan[1], Junyu Zhang[2], Jing Liu[3], Masaharu Kobayashi[4], Jixuan Wu[1, †], and Jiezhi Chen[1, †]

[1]School of Information Science and Engineering (ISE), Shandong University, Qingdao 266200, China

[2]Neumem Co., Ltd, Hefei 230093, China

[3]Key Laboratory of Microelectronic Devices and Integrated Technology, Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100084, China

[4]Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

Abstract: With the rapid development of machine learning, the demand for high-efficient computing becomes more and more urgent. To break the bottleneck of the traditional Von Neumann architecture, computing-in-memory (CIM) has attracted increasing attention in recent years. In this work, to provide a feasible CIM solution for the large-scale neural networks (NN) requiring continuous weight updating in online training, a flash-based computing-in-memory with high endurance (10^9 cycles) and ultra-fast programming speed is investigated. On the one hand, the proposed programming scheme of channel hot electron injection (CHEI) and hot hole injection (HHI) demonstrate high linearity, symmetric potentiation, and a depression process, which help to improve the training speed and accuracy. On the other hand, the low-damage programming scheme and memory window (MW) optimizations can suppress cell degradation effectively with improved computing accuracy. Even after 10^9 cycles, the leakage current (I_{off}) of cells remains sub-10pA, ensuring the large-scale computing ability of memory. Further characterizations are done on read disturb to demonstrate its robust reliabilities. By processing CIFAR-10 tasks, it is evident that ~90% accuracy can be achieved after 10^9 cycles in both ResNet50 and VGG16 NN. Our results suggest that flash-based CIM has great potential to overcome the limitations of traditional Von Neumann architectures and enable high-performance NN online training, which pave the way for further development of artificial intelligence (AI) accelerators.

Key words: NOR flash memory; computing-in-memory; endurance; neural network; online training

Citation: Y Feng, Z H Sun, Y R Qi, X P Zhan, J Y Zhang, J Liu, M Kobayashi, J X Wu, and J Z Chen, Optimized operation scheme of flash-memory-based neural network online training with ultra-high endurance[J]. *J. Semicond.*, 2024, 45(1), 012301. https://doi.org/10.1088/1674-4926/45/1/012301

1. Introduction

To address the concerns from frequent data-shuttling-related energy consumption and latency, computing-in-memory (CIM) applications in neural networks (NNs) have attracted much attention. Some previous works have demonstrated CIM-based NN inference[1–3], but it is still challenging to implement online training due to strict requirements on both performance (speed, power, array size, etc.) and reliability (endurance, stabilities, etc.). Though NNs using emerging memories, such as random-access memory (RRAM)[4] and phase change memory (PCM)[5], and have demonstrated great performance on CIM inference, memory performance for online training applications needs to be further improved. Due to excellent electrical performance such as endurance and programming speed, RRAM has emerged as one of the most promising candidates for the synapse of the NN online training[6, 7], however, the reliability and non-linearity in large arrays still need further optimizations. Flash-based CIM provides a more feasible and reliable solution because of its mature technology, ultra-high bit density, and capabilities to construct large arrays for matrix operations. So far, CIM architectures can be applied in NNs with software-combined offline training, while strict requirements have arisen for endurance and programming speed for NN online training in CIM hardware. Thus, for further exploration of flash-based CIM as online training NN accelerators, it is strongly required to break the obstacles of endurance and the speed of flash cells.

2. Background

2.1. Flash-based CIM architecture

Fig. 1 illustrates the architecture of flash memory and the adopted CIM scheme in this work. The matrix-vector-multiplication (MVM) is implemented through a matrix represented by I_d which can be tuned by V_{th}, and vectors represented by the pulse time of V_g. Firstly, the matrix and vectors are pre-processed to the corresponding electrical parameters of the device array. The matrix needs to be stored in the array and the vectors are transferred to the pulse time. Then, the amount of charge can represent the output of MVM, which can be described as $Q = I \cdot t$. The great reliability and mature technology of flash memory ensure the computing accuracy of large-scale MVM.

Correspondence to: J X Wu, jixuanwu@sdu.edu.cn; J Z Chen, chen.jiezhi@sdu.edu.cn

Received 14 JULY 2023; Revised 9 SEPTEMBER 2023.

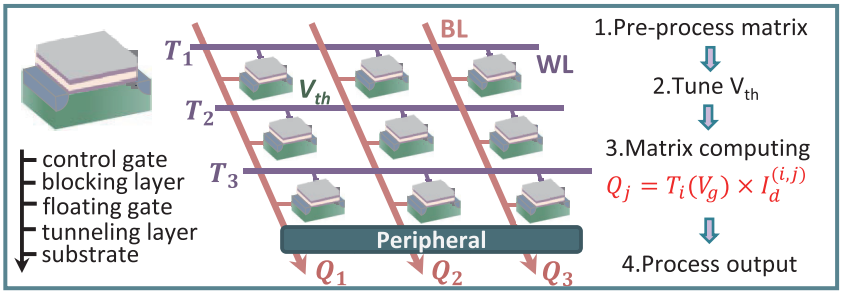


Fig. 1. (Color online) Schematics of flash-based CIM architecture. The pulse time of V_g and the threshold voltage is individually mapped as vector and matrix, then the amount of charge can represent the result of MVM.

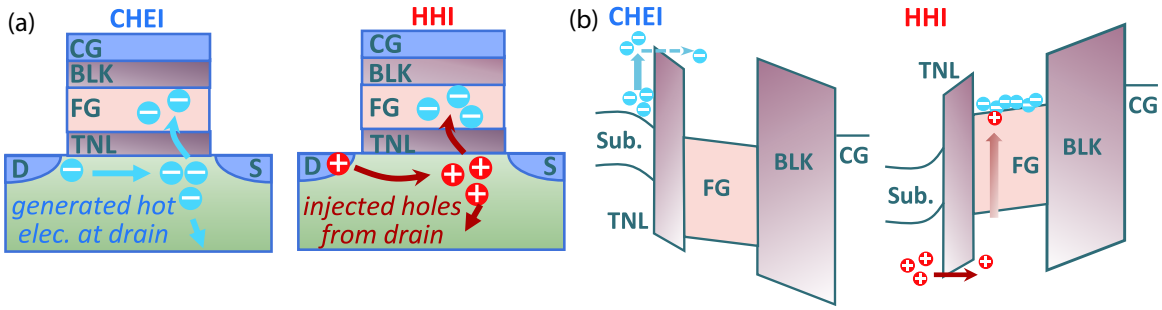


Fig. 2. (Color online) (a) Schematic of adopted CHEI and HHI programming scheme. (b) The energy band diagram of CHEI and HHI programming scheme.

2.2. Flash operation by CHEI and HHI

The CHEI programming scheme and the HHI erasing scheme are adopted in this work. The separated word line (WL) and bit line (BL) of NOR flash allow single-cell selectivity and individual programming operations with CHEI. Different from the traditional Fowler-Nordheim (FN) tunneling operation, the HHI can tune cells individually, benefitting from the independent WL and BL[8]. In addition, the applied voltage is much lower than FN tunneling. The scheme and energy band diagram of the CHEI and the HHI are shown in Fig. 2. The model of CHEI can be described as that the energetic electrons gain energy primarily from the lateral channel field (or the applied drain bias V_d) to overcome the Si–SiO_2 potential barrier, injecting into the floating gate layer to increase V_{th}. The HHI scheme utilizes the positive voltage on BL and negative voltage on WL, and the band-to-band tunneling (BTBT) is happening at the cell drain junction. Meanwhile, electrons generated by this process are gathered by the drain contact. With the large bias of the p-well and drain junction, the holes that flow towards the p-well can be injected into the floating gate under the high negative voltage of the gate. In this process, holes injected into the floating gate can recombine with electrons originally stored in the floating gate, thereby reducing the cell V_{th}.

2.3. ResNet50 and VGG16 neural networks

We choose the representative ResNet50 and VGG16 neural networks to test the performance of the proposed online training architecture. The system frame diagram of both two architectures is shown in Fig. 3. ResNet50 is a convolutional neural network architecture that was proposed to simplify the training of deeper networks and improve their speed and accuracy. The key innovation is the introduction of residual blocks, which allow the network to learn identity mappings and avoid the degradation problem caused by adding more layers. ResNet50 is representative of ResNet and has achieved a state-of-the-art performance on various image recognition tasks. On the other hand, VGG16 is another popular convolutional neural network architecture that also focuses on increasing network depth. One of its key improvements over previous architectures such as AlexNet is the use of smaller 3 × 3 convolutional filters instead of larger ones. The VGG16 model has achieved widespread adoption in multiple domains due to its ability to enable better approximation of complex functions while maintaining a manageable number of parameters. It has also been shown to achieve high accuracy on image recognition tasks.

3. Reliability analysis and discussion

The 55-nm NOR flash array is used to construct the CIM matrix for large-scale online training NNs, wherein the CHEI and the HHI[9] are adopted for ultra-fast threshold voltage (V_{th}) tuning in cells for weight updating (Fig. 4). In the conventional operation scheme towards storage usages, the MW has to be large to make the error bits as low as possible. While for NN applications with a stronger tolerance to noise, we can optimize the MW to achieve faster operation speed and suppress the cell degradation with fewer pulses and lower programming biases.

Rather than the 7.7 V substrate operation voltage and −8 to −9.5 V gate operation voltage used in FN tunneling operation, the 0 V substrate operation voltage of the HHI makes the design of the peripheral circuit much simpler. More importantly, it is found that as fast as 10 ns (pulse width) operations can be adopted for V_{th} tuning. The HHI can achieve 10^4 times faster erasing than FN tunneling as shown in Fig. 4(b). The fast programming speed is essential for the frequently fast weight updating of NN online training. Furthermore, it is necessary to ensure the linearity of the conductance–pulse curve[10] for online training applications. The pro-

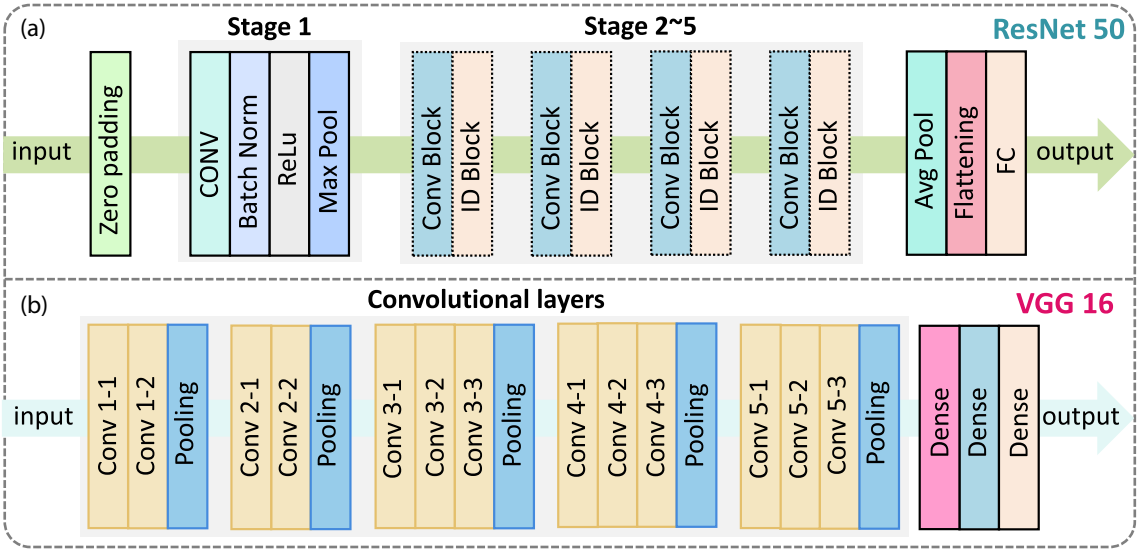


Fig. 3. (Color online) The architecture of (a) ResNet 50 and (b) VGG 16 convolutional neural network.

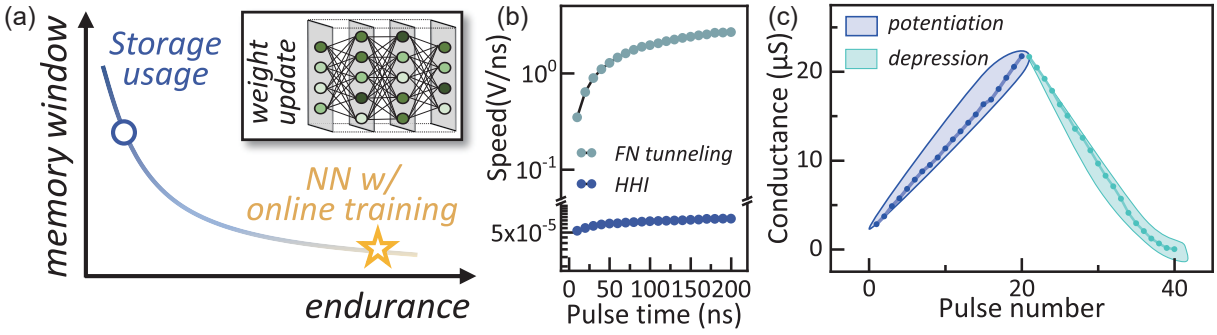


Fig. 4. (Color online) (a) The proposed scheme to improve both endurance and speed by optimizing the operation scheme for NN online training. (b) The comparison of the V_{th} tuning speed of FN tunneling and the HHI. (c) The high linearity and symmetric potentiation and depression process using the CHEI and the HHI combined methods.

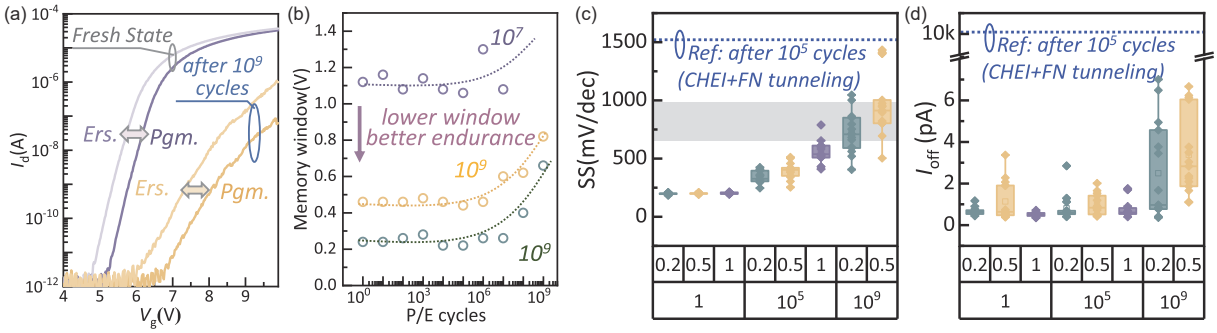


Fig. 5. (Color online) (a) The *I*–*V* curves of the programmed/erased state before and after 109 cycles. (b) Enhancements of endurance at lower MW show the trade-off between MW and endurance. (c) SS value and (d) I_{off} of different MW and cycles compared with the traditional programming scheme, wherein each box contains 15 different memory cells.

posed CHEI and HHI combined method shows high linearity in Fig. 4(c), as well as the symmetric potential and depression process during the short programming of 10 ns and erasing time.

In addition to adjustments in the programming and erasing scheme, the trade-off between MW and endurance is investigated in this work for CIM applications, especially for NN online training. Impressively, by adopting the CHEI–HHI combined programming scheme and lowering the MW, flash cells can realize record high endurance, exceeding 10^9 cycles in 0.2–0.5 V MW operations, which is enough for 1-bit/cell and even 2-bit/cell operations in CIM applications. In Fig. 5(a), the *I*–*V* curves demonstrate that even after 10^9 cycles, the current is quite stable for computing. Fig. 5(b) shows the balance between the MW and endurance. Although the test results show that the endurance decreases with the increment of the MW, it should be noted that even 4 bit/cell operations (1 V MW) can achieve 10^7 endurance by adopting the proposed programming scheme, which is enough for various large-scale NNs. However, many devices including flash memory exhibit state-dependent programming variation, and programming variations in multibit flash memories are generally

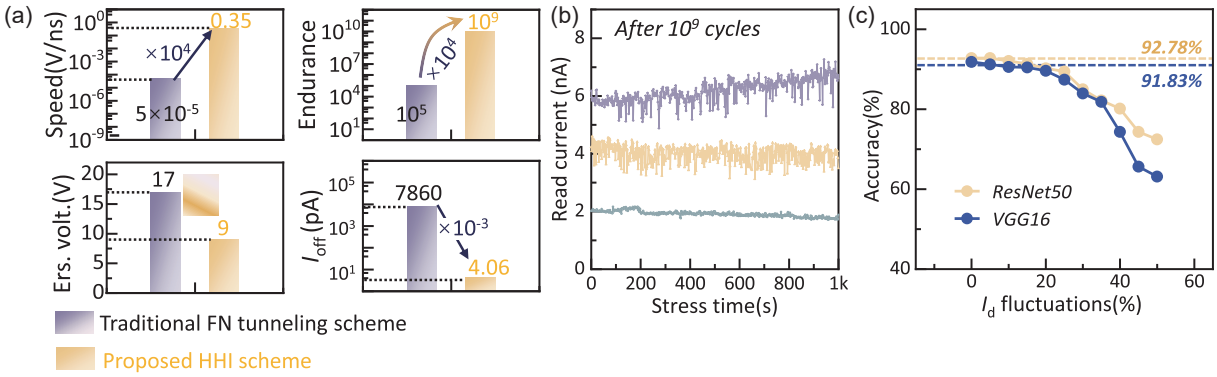


Fig. 6. (Color online) (a) Comparisons between the proposed scheme and the traditional scheme. (b) Read disturbance of different states after 10^9 cycles. (c) Applications in CIFAR-10 using ResNet50 and Vgg16. Even after 10^9 cycles, ~90% accuracy can be achieved for the CIFAR-10 task.

Table 1. The benchmark of this work and various non-volatile CIM devices.

Ref	Cell type	On/off ratio	Pgm.speed	Endurance	DR (s)
[8]	PCM	10^4	–	10^8	10^6
[13]	RRAM	10^3	1 μs	–	10^4
[14]	FeFET	10^5	300 ns	10^5	10^4
[15]	3D flash	10^5	–	10^5	10^5
[16]	Flash	10^2	10 μs	10^5	–
This work (optimized operation)	Flash	10^6 (before cycles)	10 ns	10^9	10^5

more state-independent while also suffering from additional nonlinear behaviors[11]. Therefore, compared with the larger MW with multi-level-cell (MLC) operation mode, the small MW may not cause degradation in the prediction accuracy when the MW is enough for one-bit programming precisely in the SLC operation mode.

After cycling, the subthreshold swing (SS) and the leakage current (I_{off}) are also tested to avoid deteriorated computing accuracy as the result of on/off ratio degradation. Although both the traditional FN tunneling and the HHI will increase the SS value, the increment of SS is conducive to precise device programming for more precise weight updating of the CIM application instead. This is because larger SS values result in larger memory windows between adjacent programming states. Therefore, only the degradation of the I_{off} will significantly impact the computational performance to the deteriorated on/off ratio. However, different from FN tunneling erasing with serious I_{off} degradation, I_{off} can be suppressed to sub-10 pA after 10^9 cycles by the HHI erasing (Fig. 5(d)). This can be understood because FN tunneling causes serious degradation to gate dielectrics and the interface, while the HHI mainly degrades the interface[12].

This can be observed from the statistical data of SS and I_{off} in Fig. 5. After 10^9 cycles, the subthreshold region of cells with larger SS and sub-10pA I_{off} allows a larger V_{th} tuning range and better immunity to the fluctuations, which can be utilized to construct efficient large-scale online training NNs.

4. Performance in neural network

The comparison is analyzed between different programming schemes. Aiming at CIM applications, the test results of 55 nm flash memory in Fig. 6(a) show that the traditional FN tunneling scheme is much slower and requires a higher tuning voltage as compared to the proposed HHI scheme. Besides this, the I_{off} after P/E cycles increase significantly, which can have a considerable impact on the on/off ratio of the device, ultimately affecting the CIM accuracy.

The read disturb (RD) characteristic is then tested to evidence robust reliabilities in flash cells. Well-controlled RD lasting for 1 ks can be observed in Fig. 6(b). After 10^9 cycles, though the read current decreases to nA-level as a result of the increment of SS, the RD is quite stable. The long-term read disturbance characteristics have an ignorance impact on computing accuracy with cells' current fluctuations for online training of NN. The short-term currency fluctuations will not degrade the accuracy of neural network calculations.

To further evaluate the performance of the compact flash-based NN system, the chip tester is designed to characterize flash-based CIMs and can support fast programming, as well as operations of matrix-vector multiplication (MVM). To demonstrate the feasibility of flash CIM, standard the ResNet50 and VGG16 convolutional neural network (CNN) is implemented for image classification in the CIFAR-10 dataset with 10 object classes, as shown in Fig. 6(c). With the simulations of different NNs, the high accuracy of the proposed system is demonstrated. It should be noted that short-term RDs are an important impact factor because online training requires continuous ultra-fast weight updating. Impressively, over 90% recognition accuracy has been achieved after 10^9 cycles with the proposed flash CIM. The comparisons with other related works are summarized in Table 1. The results indicate that the proposed programming scheme for NOR flash demonstrates superior characteristics in terms of on/off ratio, programming speed, endurance, and DR compared to other emerging memories.

5. Conclusion

This work shows the potential of flash-based computational-in-memory (CIM) devices to achieve high endurance (10^9) and ultra-fast programming speed (10 ns) through the implementation of the CHEI–HHI programming scheme and MW optimizations. Utilizing this optimized operation scheme,

a compact flash-based NN online training CIM system is proposed. Our results demonstrate that even after 10^9 cycles, a high accuracy rate of approximately 90% can be attained when performing CIFAR-10 tasks. Further characterizations are done on read disturb to evidence robust reliabilities, highlighting its significant potential for online training of actual NN tasks. This work provides a comprehensive assessment of a flash-based online training network with potential implications for the advancement of AI accelerators.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62034006, 92264201, and 91964105), the Natural Science Foundation of Shandong Province (Nos. ZR2020JQ28 and ZR2020KF016), and the Program of Qilu Young Scholars of Shandong University.

References

[1] Yao P, Wu H Q, Gao B, et al. Fully hardware-implemented memristor convolutional neural network. Nature, 2020, 577, 641

[2] Khwa W S, Akarvardar K, Chen Y S, et al. MLC PCM techniques to improve nerual network inference retention time by 105X and reduce accuracy degradation by 10.8X. Proc IEEE Symp VLSI Technol, 2020, 1

[3] Zhang W Y, Wang S C, Li Y, et al. Few-shot graph learning with robust and energy-efficient memory-augmented graph neural network (MAGNN) based on homogeneous computing-in-memory. 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2022, 224

[4] Kumar S, Wang X X, Strachan J P, et al. Dynamical memristors for higher-complexity neuromorphic computing. Nat Rev Mater, 2022, 7, 575

[5] Lu Y M, Li X, Yan B N, et al. In-memory realization of eligibility traces based on conductance drift of phase change memory for energy-efficient reinforcement learning. Adv Mater, 2022, 34, 2107811

[6] Huang P, Zhou Z, Zhang Y, et al. Dual-configuration in-memory computing bitcells using SiO_x RRAM for binary neural networks. APL Mater, 2019, 7, 081105

[7] Chang C C, Chen P C, Chou T, et al. Mitigating asymmetric nonlinear weight update effects in hardware neural network based on analog resistive synapse. IEEE J Emerg Sel Top Circuits Syst, 2018, 8, 116

[8] Ravsher T, Garbin D, Fantini A, et al. Enhanced performance and low-power capability of SiGeAsSe-GeSbTe 1S1R phase-change memory operated in bipolar mode. 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2022, 312

[9] Ielmini D, Ghetti A, Spinelli A S, et al. A study of hot-hole injection during programming drain disturb in flash memories. IEEE Trans Electron Devices, 2006, 53, 668

[10] Wu W, Wu H Q, Gao B, et al. A methodology to improve linearity of analog RRAM for neuromorphic computing. 2018 IEEE Symposium on VLSI Technology, 2018, 103

[11] Wang Q W, Park Y, Lu W D. Device variation effects on neural network inference accuracy in analog In-memory computing systems. Adv Intell Syst, 2022, 4, 2100199

[12] Ogawa S, Shiono N. Interface-trap generation induced by hot-hole injection at the $Si\text{-}SiO_2$ interface. Appl Phys Lett, 1992, 61, 807

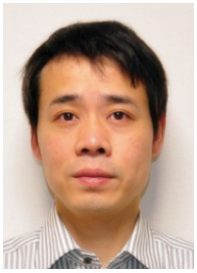
[13] Choi W, Kwak M, Heo S, et al. Hardware neural network using hybrid synapses via transfer learning: WO_x nano-resistors and TiO_x RRAM synapse for energy-efficient edge-AI sensor. 2021 IEEE International Electron Devices Meeting (IEDM), 2021, 23.1. 1

[14] Ali T, Seidel K, Kühnel K, et al. A novel dual ferroelectric layer based MFMFIS FeFET with optimal stack tuning toward low power and high-speed NVM for neuromorphic applications. 2020 IEEE Symposium on VLSI Technology, 2020, 1

[15] Lue H T, Hsu P K, Wei M L, et al. Optimal design methods to transform 3D NAND flash into a high-density, high-bandwidth and low-power nonvolatile computing in memory (nvCIM) accelerator for deep-learning neural networks (DNN). 2019 IEEE International Electron Devices Meeting (IEDM), 2020, 38.1.1

[16] Malavena G, Spinelli A S, Compagnoni C M. Implementing spike-timing-dependent plasticity and unsupervised learning in a mainstream NOR flash memory array. 2018 IEEE International Electron Devices Meeting (IEDM), 2019, 2.3.1

Yang Feng received a BEng degree from the School of Information Science and Engineering (ISE), Shandong University, in 2021, where she is currently pursuing a PhD degree with the School of Information Science and Engineering (ISE), Shandong University. Her focuses are on the design and simulation of computing-in-memory circuits and systems.



Jiezhi Chen received a PhD degree from the Department of Informatics and Electronics, The University of Tokyo, in 2009. In 2010, he joined the Research and Development Center, Toshiba Corporation. He is currently a professor with the School of Information Science and Engineering, Shandong University, China. His research interests include the characterization and process engineering of nano-scale transistors and non-volatile memories, with a main focus on reliability physics.